Paper presented at HPCMP Users Group Conference
Seattle, Washington
14-17 July 2008

PIPA: A High-Throughput Pipeline for Protein Function Annotation

Chenggang Yu, Valmik Desai, Nela Zavaljevski, and Jaques Reifman
Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD 21702, USA

{cyu, valmik, nelaz}@bioanalysis.org, jaques.reifman@us.army.mil

Abstract

Traditional experimental methods to determine the functions of proteins encoded in genomic sequences cannot keep pace with the avalanche of sequence data produced by new high-throughput sequencing technologies. This prompted the development of numerous bioinformatics approaches for automated protein function annotation. However, different function classification terminologies are frequently used by these different approaches, precluding the integration of multisource predictions. We developed Pipeline for Protein Annotation (PIPA), a genome-wide protein function annotation pipeline that runs in a high-performance computing environment. PIPA integrates different tools and employs the Gene Ontology (GO) to provide consistent annotation and resolve prediction conflicts.

PIPA has three modules that allow for easy development of specialized databases and integration of various bioinformatics tools. The first module, the pipeline execution module, consists of programs that enable the user access to and control of the pipeline's parallel execution of multiple jobs, each searching a particular database for a chunk of the input data. The execution module wraps the second module, the core pipeline module. The integrated resources, the program for terminology conversion to GO, and the consensus annotation program constitute the main components of the core module. The third module is the preprocessing module. This last module contains the program for customized generation of protein function databases and the GO-mapping generation program, which creates GO mappings for the terminology conversion program.

The current implementation of PIPA annotates protein functions by combining the results of an in-house-developed database for enzyme catalytic function prediction (CatFam) and the results of multiple integrated resources, such as the 11 member databases of InterPro and the Conserved Domains Database, into common GO terms. A Web-page-based graphical user interface is developed based on the User Interface Toolkit. The pipeline is deployed on two LINUX clusters, JVN at the Army Research Laboratory Major Shared Resource Center and JAWS at the Maui High Performance Computing Center. Currently, scientists at the Naval Medical Research Center are using PIPA to predict protein functions for newly sequenced bacterial pathogens and their near-neighbor strains.

Validation tests show that, on average, the CatFam database yields predictions of enzyme catalytic functions with accuracy greater than 95%. Test results of the consensus GO annotation show an improvement in performance of up to 8% when compared with annotations in which consensus is not used.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
PIPA: A High-Throughput Pipeline for Protein Function Annotation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,Biotechnology HPC Software Applications Institute,Telemedicine and Advanced Technology Research Center,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Traditional experimental methods to determine the functions of proteins encoded in genomic sequences cannot keep pace with the avalanche of sequence data produced by new high-throughput sequencing technologies. This prompted the development of numerous bioinformatics approaches for automated protein function annotation. However, different function classification terminologies are frequently used by these different approaches, precluding the integration of multisource predictions. We developed Pipeline for Protein Annotation (PIPA), a genome-wide protein function annotation pipeline that runs in a high-performance computing environment. PIPA integrates different tools and employs the Gene Ontology (GO) to provide consistent annotation and resolve prediction conflicts. PIPA has three modules that allow for easy development of specialized databases and integration of various bioinformatics tools. The first module, the pipeline execution module, consists of programs that enable the user access to and control of the pipeline?s parallel execution of multiple jobs, each searching a particular database for a chunk of the input data. The execution module wraps the second module, the core pipeline module. The integrated resources, the program for terminology conversion to GO, and the consensus annotation program constitute the main components of the core module. The third module is the preprocessing module. This last module contains the program for customized generation of protein function databases and the GO-mapping generation program, which creates GO mappings for the terminology conversion program. The current implementation of PIPA annotates protein functions by combining the results of an in-house-developed database for enzyme catalytic function prediction (CatFam) and the results of multiple integrated resources, such as the 11 member databases of InterPro and the Conserved Domains Database, into common GO terms. A Web-page-based graphical user interface is developed based on the User Interface Toolkit. The pipeline is deployed on two LINUX clusters, JVN at the Army Research Laboratory Major Shared Resource Center and JAWS at the Maui High Performance Computing Center. Currently, scientists at the Naval Medical Research Center are using PIPA to predict protein functions for newly sequenced bacterial pathogens and their near-neighbor strains. Validation tests show that, on average, the CatFam database yields predictions of enzyme catalytic functions with accuracy greater than 95%. Test results of the consensus GO annotation show an improvement in performance of up to 8% when compared with annotations in which consensus is not used.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

1. Introduction

As advanced high-throughput sequencing technologies accumulate genome sequences at an ever-increasing rate,[1] computational methods become mandatory to annotate them. Basic annotation entails search for the genomic regions that code for proteins or RNA, transcription factors, insertion elements, sequence repeats, and other genomic elements.[2] Identified protein coding regions are then annotated using protein function prediction methods, ranging from *ab initio* to genomic context based to sequence based .[3]

Function prediction based on sequence similarity is the most widely used computational approach. The underlying assumption is that proteins with similar sequences share similar functions. The BLAST program[4] is usually used to search for similar sequences in large databases. Compared to direct sequence search methods like BLAST, predictions based on function-related sequence features, such as sequence domains or motifs, are more accurate and more sensitive, in particular for proteins that have low sequence similarity with proteins of known function. This has led to the development of a wide variety of general-purpose feature databases, such as Pfam,[5] the Clusters of Orthologous Groups (COG),[6] and the Conserved Domains Database (CDD).[7] Recently, customized feature databases have been developed for the prediction of specific protein functions. For example, PRIAM[8] is a specialized database for protein catalytic function predictions, which has proven to be more accurate and more sensitive than feature databases developed for general-purpose protein function predictions.

With the existence of many programs and databases that infer different protein functions, large integrated information systems, such as InterPro[9] and IMG,[10] have been developed. These systems include comprehensive resources that allow curators and users alike to gain insights into protein functions. However, these systems are not designed to algorithmically combine different resources for automated protein function prediction. Rather, function information from different resources is usually listed in their original forms, such as accession numbers in a database, and the succinct description of protein functions, reconciling the results from the different resources and eliminating false positive predictions, is edited by human curators. A more useful application may be achieved if these resources are integrated into a protein function annotation pipeline that can exploit the specific advantages of each resource. However, different function classification terminologies are frequently used and it becomes difficult to fuse multiple predictions. In addition, the computation time needs to be considered when many time-consuming methods are integrated.

To address these issues we developed Pipeline for Protein Annotation (PIPA),[11] a genome-wide protein function annotation pipeline that runs in a high-performance computing environment. PIPA's applications range from helping address fundamental questions, such as the analysis of protein function diversity and evolution in the microbial world, to more direct biodefense-related applications that compare the functional repertoire of pathogenic and nonpathogenic organisms for improved diagnostics. Currently, scientists at the Naval Medical Research Center are using PIPA to predict protein functions for newly sequenced bacterial pathogens and their near-neighbor strains.

2. Methods and Implementation

PIPA differs from other integrated systems because it not only integrates existing programs and databases, but it also allows integration of users' data to predict particular protein functions. This is accomplished through a customized database generation procedure for user-categorized protein functions. In the generated database, proteins of the same function are grouped based on their sequence features in such a way that the proteins in each group pass predetermined thresholds. Such thresholds ensure that a database search for new proteins with unknown function yields predictions that satisfy a pre-specified nominal false positive rate. This strategy reduces error propagation, which is of concern in automated generation of protein annotation databases.

Most importantly, PIPA integrates different protein function prediction resources into a consistent and parsimonious consensus function annotation, a valuable feature that most integrated systems do not provide. Due to the different terminologies used by different inference methods, it is challenging for automated computer programs to perform consensus annotation. PIPA uses the Gene Ontology[12] (GO), which is becoming a function annotation standard in the bioinformatics community, as a unifying terminology. Hence, to map predictions to GO for resources that employ a different terminology, we developed an algorithm that automatically generates such mappings. PIPA performs consensus annotation through a novel algorithm that takes into account the hierarchical structure of GO. In a recent publication,[11] we provided a detailed description of PIPA's main algorithms.

PIPA has a modular architecture that allows for easy development of specialized databases and integration of different inference methods and databases. Figure 1 shows the three main modules of the PIPA pipeline. The first module, the pipeline execution module, consists of programs that enable the user access to and control of the pipeline's parallel execution of multiple jobs, each searching a particular database for a chunk of the input data. This module wraps the second module, the core pipeline module. The integrated resources, the program for terminology conversion to GO, and the consensus annotation program constitute the main components of the core module. The third module, the preprocessing module, contains the program for customized generation of protein function databases and the GO-mapping generation program, which creates GO mappings for the terminology conversion program.

Currently, the major methods and databases that have been integrated in PIPA consist of the in-house-developed enzyme catalytic function prediction database, CatFam, and 15 publicly available databases and resources, including InterPro member databases and the CDD. A complete list of these resources is provided by Yu *et al.* (2008).[11] PIPA takes as input protein sequences in FASTA (http://www.ebi.ac.uk/help/formats_frame.html) or GenBank[13] format and executes all integrated methods in parallel. The user specifies input parameters through a Web-page-based graphical user interface (GUI) developed using the User Interface Toolkit. Parameters recommended by the developers of each of the integrated methods are provided as the default settings. However, the parameters can be modified to control the rate of false positive predictions. These predictions, based on their original terminologies, are converted into GO terms using mapping files. Lastly, the GO consensus annotation algorithm takes these GO terms and infers consensus terms that are saved, together with the original predictions, in an output file in the General Feature Format (http://www.sanger.ac.uk/Software/formats/GFF/). PIPA's outputs are also presented through a GUI (Figure 2), which contains hyperlinks to the original Web sites of the integrated programs. The PIPA pipeline is deployed on two LINUX clusters, JVN at the Army Research Laboratory Major Shared Resource Center and JAWS at the Maui High Performance Computing Center.

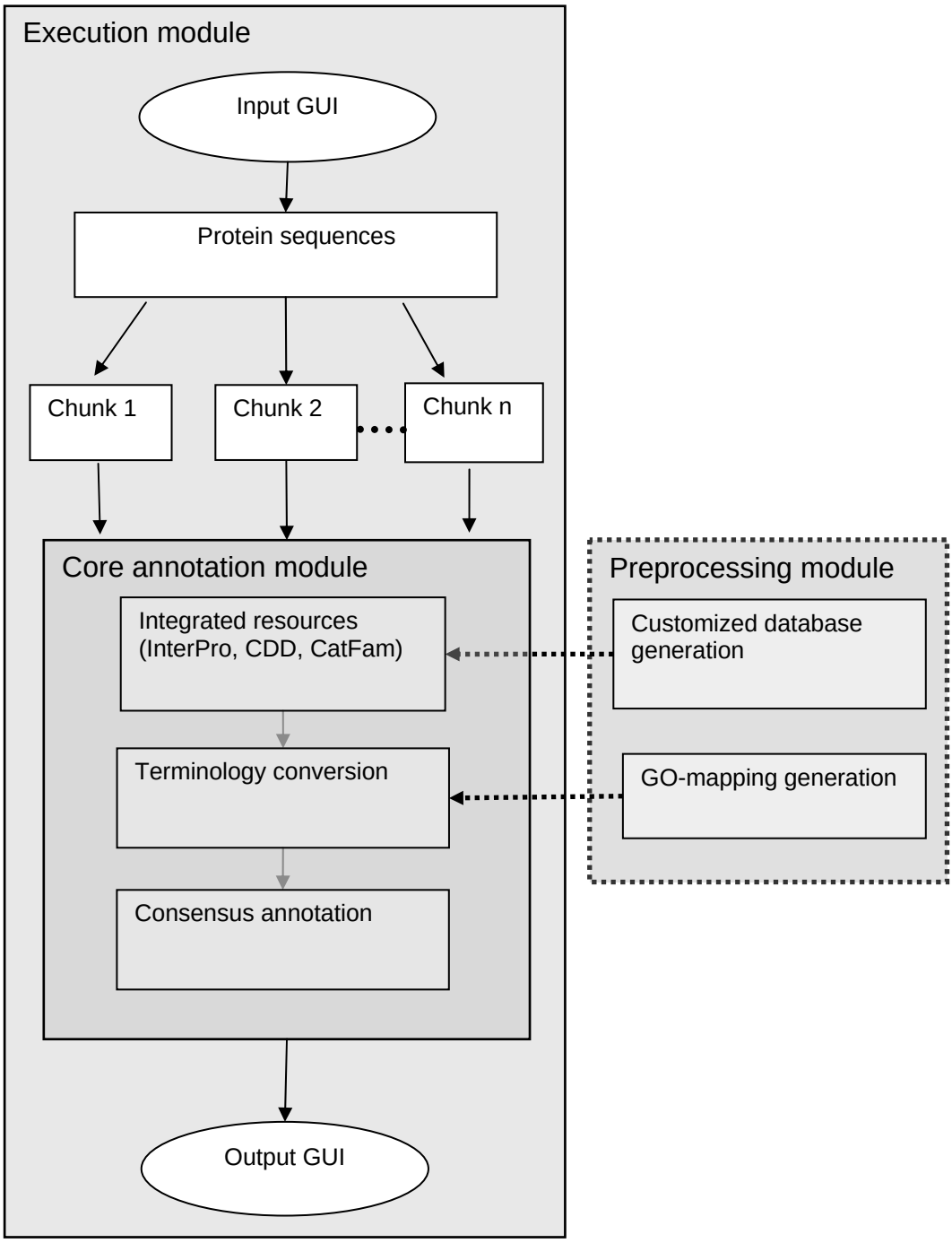

Figure 1. PIPA's key modules. PIPA's programs are organized into three modules. The pipeline execution module consists of programs that enable user access to and control of the pipeline's parallel execution of multiple programs. The execution module wraps the core module, containing the integrated resources, the terminology conversion program, and the consensus annotation program. The preprocessing module contains the customized database generation program, which is used to generate CatFam, and the GO-mapping generation program.

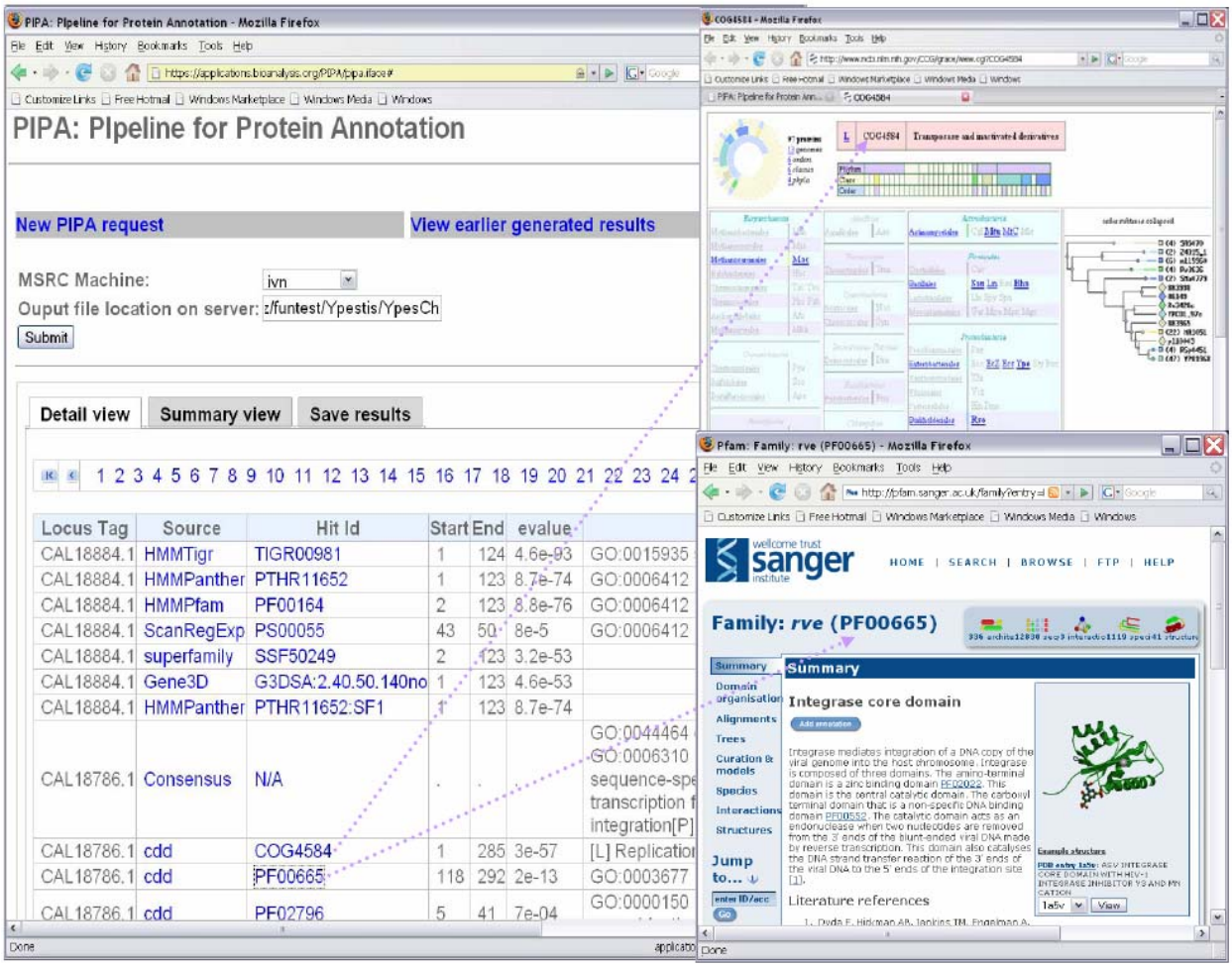

Figure 2. PIPA's output GUI, showing one of the 1174 pages of protein annotations for *Yersinia pestis* CO92. The predictions based on the integrated methods are hyperlinked if their Web sites are available. Two of the integrated methods, Pfam and COG, are indicated in the figure.

3. Results

PIPA is particularly suited for whole-genome function annotation of bacterial proteins. On the JAWS cluster, a typical bacterial genome with about 4,000 protein coding regions (e.g., *Yersinia pestis*, the causative agent of plague and a potential bio-weapon) can be annotated using all of the 16 integrated programs within 2 hr and 44 min on 64 processors. Annotation results for this genome can be obtained in 1 hr and 16 min using 128 processors; however, as shown in Figure 3, the rate of parallelization speedup (i.e., parallelization efficiency) decreases as the number of processors increases. As a rule of thumb, up to 150 processors are recommended if the user needs fast throughput.

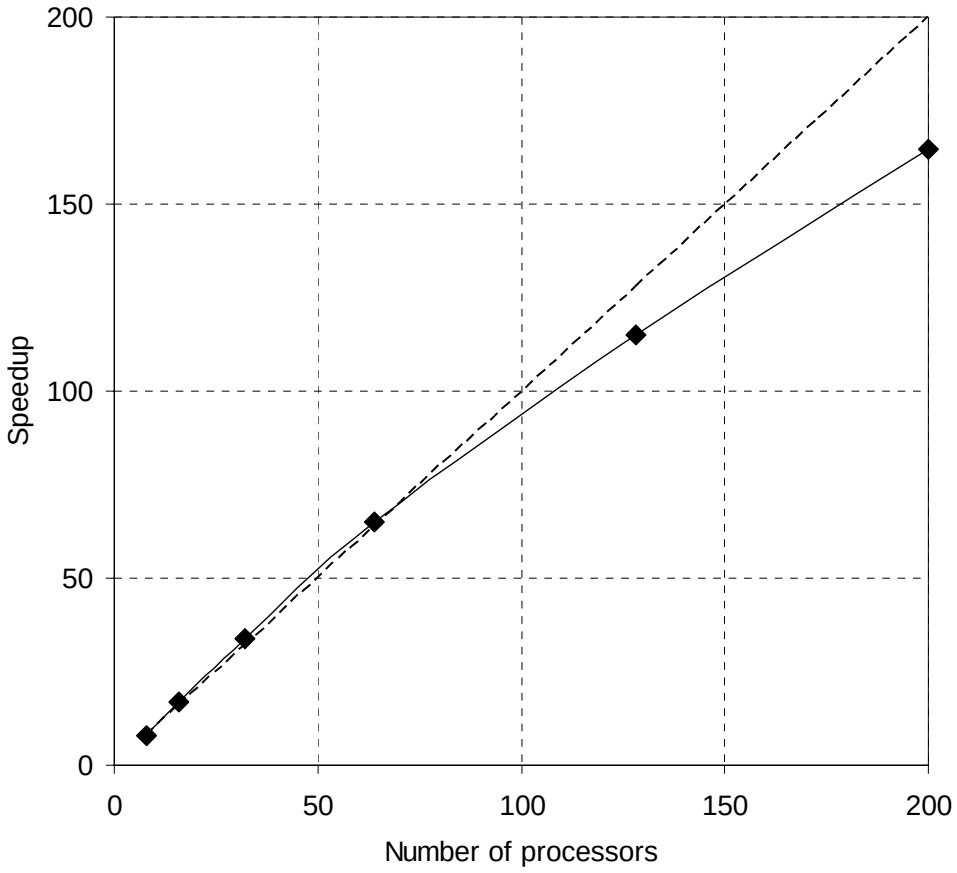

Figure 3. PIPA's parallelization efficiency for annotation of a typical bacterial genome using the JAWS cluster. All of the 16 PIPA resources are used for annotation. The speedup rate per processor (efficiency) deteriorates after 128 processors.

Among the various protein functions to be annotated, enzyme catalytic functions are of great importance. About 30% of the genes in bacterial genomes code for enzymes,[14] which play many critical roles in a variety of biological processes.[15] To enable whole-genome annotation of enzymes, we used PIPA's customized database generation algorithm and developed a version of the CatFam database with a nominal false positive rate of 10%. This setting enables a good trade-off between

accuracy and coverage. We validated the CatFam's predictions using a testing set of nearly 20,000 proteins (both enzymes and nonenzymes) not included in the database generation process and compared them with those of a similar well-established database, PRIAM. Throughout the paper, we use precision as a measure of prediction accuracy and recall as a measure of prediction coverage. Precision is the fraction of function predictions of a particular method that agrees with the gold standard annotations, while recall is the fraction of the gold standard function annotations that are predicted by a particular method. For this test, CatFam achieves a precision of 95.9% and recall of 97.0%, compared with PRIAM's precision of 82.6% and recall of 87.9%.

We also used CatFam to predict catalytic functions for 13 bacterial genomes of biodefense interest listed in Table 1, including 11 category A and B bacterial pathogens listed by the Centers for Disease Control and Prevention. For benchmarking purposes, we consider the enzyme annotations in the KEGG database (http://www.genome.jp/kegg/) as the gold standard, since these annotations combine the results of multiple resources and are partially curated. Table 1 compares the CatFam results with those obtained with PRIAM. The comparisons indicate that the CatFam predictions yield larger precision than those of PRIAM for all 13 genomes. CatFam's precision for each of the genomes is in the 70-80% range, except for the recently sequenced *Clostridiun botulinum*, which has only 21 proteins recorded in the manually annotated protein database Swiss-Prot.[16] However, CatFam's recall in three cases is substantially lower than that of PRIAM. This is consistent with the fact that PRIAM often predicts more enzymes than CatFam, increasing recall at the expense of deteriorating precision. Compared with PRIAM, CatFam is a more conservative tool, optimized for accurate enzyme function predictions.

Table 1. Whole-genome enzyme annotation for 13 bacterial genomes, assuming KEGG as the gold standard.

Genome	Precision[a]		Recall[b]	
	CatFam	PRIAM	CatFam	PRIAM
Yersinia pestis CO92[*]	0.80	0.64	0.80	0.79
Y. pestis Microtus	0.80	0.64	0.81	0.80
Y. pseudotuberculosis IP 32953	0.79	0.63	0.79	0.78
Bacillus anthracis Ames Ancestor[*]	0.71	0.54	0.68	0.74
Brucella mallei ATCC 23344[*]	0.70	0.56	0.55	0.74
B. melitensis 16M[*]	0.81	0.62	0.57	0.68
Burkholderia pseudomallei K96243[*]	0.76	0.54	0.62	0.74
Clostridium botulinum Hall[*]	0.54	0.41	0.79	0.86
Coxiella burnetii RSA 493[*]	0.81	0.65	0.79	0.78
Francisella tularensis SCHU S4[*]	0.83	0.71	0.73	0.74
Rikettsia prowazekii Madrid E[*]	0.82	0.72	0.76	0.71
Salmonella enterica Typhi CT18[*]	0.82	0.65	0.83	0.78
Vibrio cholerae N16961[*]	0.81	0.66	0.80	0.78

[*] Centers for Disease Control and Prevention category A and B pathogens

[a] Precision=TP/(TP + FP) (measure of accuracy)

[b] Recall=TP/(TP + FN) (measure of coverage)

TP: true positives; FP: false positives; FN: false negatives

Although PIPA achieves very good performance for catalytic function annotation with the CatFam database, its performance for other categorical functions is dependent on the various integrated resources. To evaluate the performance of PIPA's consensus prediction, we employ the 31,589 proteins with annotated GO terms from the Swiss-Prot database. Performance is assessed using recall and precision evaluated in a hierarchical context.[17] Figure 4 compares the performance of GO annotations with and without the consensus algorithm for the GO molecular function category. The data points corresponding to the consensus algorithm are obtained by changing the parameters of the algorithm, which are described by Yu *et al*. (2008),[11] while the performance of GO annotations without the consensus algorithm is achieved by changing the cut-off thresholds of the integrated databases. The figure suggests a significant trade-off between precision and recall. However, for a given recall, the application of the consensus algorithm yields higher precision than when consensus is not used. Precision of molecular function predictions is improved by up to 8.0%.

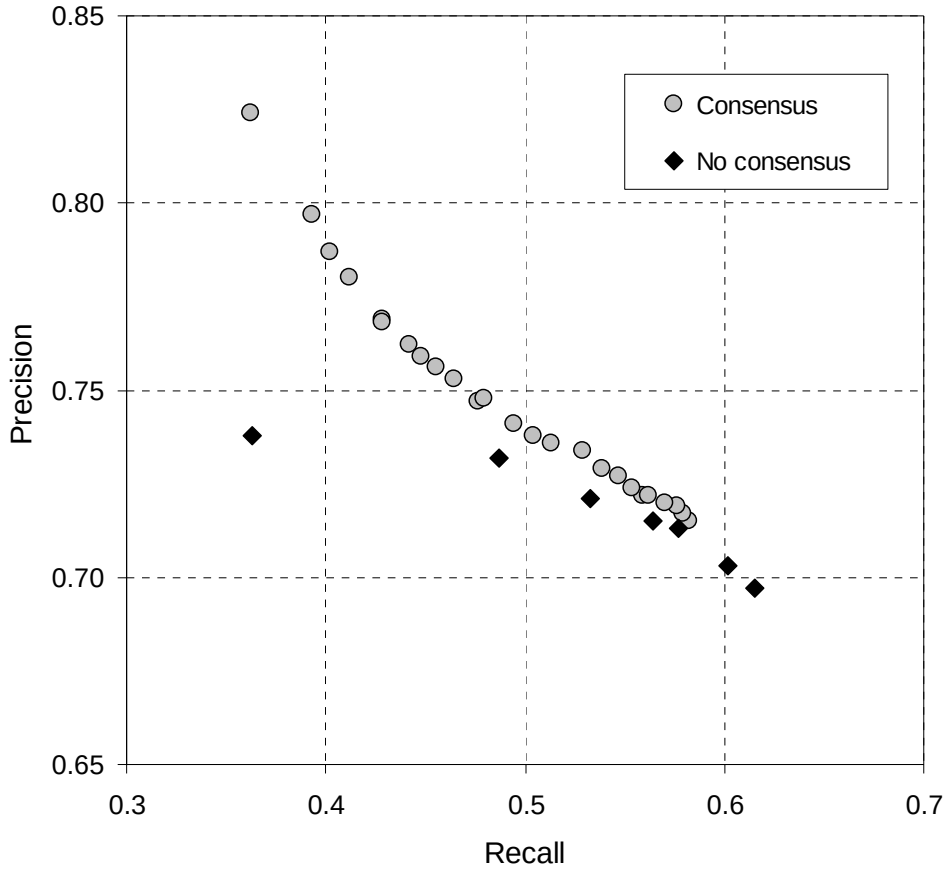

Figure 4. GO consensus evaluation. Comparison of precision and recall, evaluated using GO's hierarchical structure, for GO molecular function annotations with and without consensus. The comparison is based on 31,589 manually annotated proteins.

The results suggest that the consensus algorithm effectively integrates different function inferences to improve the precision of GO annotations. The low recall, which indicates a low coverage of GO terms predicted by the pipeline, is likely due to the incompleteness of the GO mappings that link individual databases with GO terms.

4. Conclusions

We presented PIPA, an integrated and automated protein function annotation pipeline. PIPA improves annotation accuracy by providing the means to develop customized databases and by exploiting and consistently consolidating protein function information from disparate sources based on different terminologies. An added benefit is that the consolidated function predictions are given in GO terms, which is becoming the *de facto* standard in the community.

We used PIPA's customized database generation algorithm to construct the database for catalytic function prediction, CatFam. Comparisons with a well-established resource, PRIAM, demonstrate the effectiveness of the enzyme database generation method and the CatFam database. Comparisons based on a testing dataset of 20,000 proteins and 13 bacterial genomes indicate that CatFam outperforms PRIAM in precision and, in most cases, in recall as well.

Concise and more accurate GO annotations can be obtained by the proposed consensus algorithm. The ability to optimize the algorithm's parameters and the future availability of additional reliable GO mappings will further improve PIPA's performance in whole-genome protein annotation of newly sequenced bacteria of interest for biodefense and other applications.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

Acknowledgement

This work was sponsored by the U.S. Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes initiative.

References

1. Hall N. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 2007;210(Pt 9):1518-1525.
2. Medigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 2007;158(10):724-736.
3. Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinform 2006;7(3):225-242.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-410.
5. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34(Database issue):D247-251.
6. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28(1):33-36.
7. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH,

Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 2005;33(Database issue):D192-196.

8. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 2003;31(22):6633-6639.
9. Mulder NJ, Apweiler R. The InterPro database and tools for protein domain analysis. Curr Protoc Bioinformatics 2008;Chapter 2:Unit 2 7.
10. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC. The integrated microbial genomes (IMG) system. Nucleic Acids Res 2006;34(Database issue):D344-348.
11. Yu C, Zavaljevski N, Desai V, Johnson S, Stevens FJ, Reifman J. The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. BMC Bioinformatics 2008;9:52.
12. Hill DP, Smith B, McAndrews-Hill MS, Blake JA. Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics 2008;9 Suppl 5:S2.
13. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2008;36(Database issue):D25-30.
14. Freilich S, Spriggs RV, George RA, Al-Lazikani B, Swindells M, Thornton JM. The complement of enzymatic sets in different species. J Mol Biol 2005;349(4):745-763.
15. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res 2008;36(Database issue):D480-484.
16. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol Biol 2007;406:89-112.
17. Verspoor K, Cohn J, Mniszewski S, Joslyn C. A categorization approach to automated ontological function annotation. Protein Sci 2006;15(6):1544-1549.